Markov Chain Truncation for Doubly-Intractable Inference

Markov Chain Truncation for Doubly-Intractable Inference

Colin Wei
Stanford University

Iain Murray
University of Edinburgh

Abstract

Computing *partition functions*, the normalizing constants of probability distributions, is often hard. Variants of importance sampling give unbiased estimates of a normalizer Z, however, unbiased estimates of the reciprocal $1/Z$ are harder to obtain. Unbiased estimates of $1/Z$ allow Markov chain Monte Carlo sampling of "doubly-intractable" distributions, such as the parameter posterior for Markov Random Fields or Exponential Random Graphs. We demonstrate how to construct unbiased estimates for $1/Z$ given access to black-box importance sampling estimators for Z. We adapt recent work on random series truncation and Markov chain coupling, producing estimators with lower variance and a higher percentage of positive estimates than before. Our debiasing algorithms are simple to implement, and have some theoretical and empirical advantages over existing methods.

1 Introduction

Markov Chain Monte Carlo (MCMC) algorithms can asymptotically draw samples from distributions with intractable normalizing constants. However, sampling from "doubly-intractable" distributions (Murray et al., 2006) is more challenging: direct application of MCMC methods requires the computation of an intractable normalizing constant $Z(\theta)$ at each step (Section 2.2 has an example). Until recently, the only valid MCMC methods for doubly-intractable distributions required exact samples from distributions with the relevant normalizing constants (Møller et al., 2006; Murray et al., 2006). Drawing exact samples is possible for some high-dimensional distributions (Propp and Wilson, 1998), but is hard in general.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

Lyne et al. (2015) provided the first practical and asymptotically correct MCMC method for doubly-intractable distributions that doesn't require exact sampling. This work constructs unbiased estimates of the reciprocal normalizing constants $1/Z(\theta)$ using unbiased estimates of $Z(\theta)$ obtained by importance sampling. A "Russian roulette" random series truncation debiases the estimator for $1/Z(\theta)$. The pseudo-marginal framework (Andrieu and Roberts, 2009) is then adapted to use these estimates to form an MCMC method.

Inspired by the approach of Glynn et al. (2014), we construct unbiased estimates of reciprocal normalizing constants by applying Russian roulette truncations to a Markov chain rather than an importance sampler. Swapping to Markov chains improves two aspects of the estimators, both theoretically and empirically.

First, Russian roulette estimates of the reciprocal normalizer are not guaranteed to be positive. It can be shown that there is no general procedure to construct a strictly positive unbiased estimator by debiasing estimates of the normalizer (Jacob et al., 2015). However, we find Markov chain-based estimators are positive more often than corresponding importance sampling estimators, and we test the impact of this difference on a doubly-intractable Markov chain empirically.

Second, Russian roulette forms estimates by truncating an infinite series. In the original scheme, each subsequent term in the series was estimated with an exponentially growing number of importance samples, yet it is still hard to prove that the estimator has finite expectation. Our estimator has provably finite expectation, and only requires a number of Monte Carlo samples linear in the length of the truncated series.

2 Preliminaries

For the remainder of this paper, we will assume that we are interested in the partition function $Z(\theta)$ of distributions $p(x \mid \theta) = \frac{p^*(x \mid \theta)}{Z(\theta)}$, parameterized by θ. Here, $p^*(x \mid \theta)$ is the unnormalized probability, defining the partition function $Z(\theta) = \int p^*(x \mid \theta) \, \mathrm{d}x$. We will omit the parameters θ when we only need to consider one normalizing constant.

Importance sampling can give an unbiased estimator of a normalizer Z. The method needs a target distribution $P(X)=P^*(X)/Z$ that has the normalization constant we are interested in, and a proposal distribution Q with support on the same state space $\mathcal{X}$. The unbiased estimator for Z is an average of importance weights, $w(X)=P^*(X)/Q(X)$, for states sampled from Q:

$$\mathbb{E}_{X\sim Q}\left[\frac{P^*(X)}{Q(X)}\right] = Z.$$

In general, the importance sampling target P and our original distribution of interest p do not have to be the same. For example, annealed importance sampling (AIS) (Neal, 2001), performs importance sampling on an augmented state space.

We will require an unbiased estimate of $1/Z$. Jensen's inequality states that the reciprocal of an importance sampling estimate is biased, and so needs correcting.

2.1 Russian Roulette Truncation

Russian roulette truncation can be used to obtain unbiased estimates of $1/Z$. The method was first introduced in the physics literature (Carter and Cashwell, 1975; Lux and Koblinger, 1991), while we rely on the formulation presented by McLeish et al. (2011), Glynn et al. (2014) and Lyne et al. (2015). For our specific setting, the truncation scheme depends on a sequence of estimators $Y = (Y^{(i)} : i \geq 0)$ which satisfy the property that $\lim_{i\to\infty}\mathbb{E}[Y^{(i)}] = 1/Z$. The procedure involves drawing a random integer N, independent of Y, and then taking the sum

$$S = Y^{(0)} + \sum_{i=1}^{N}\frac{Y^{(i)} - Y^{(i-1)}}{\Pr(N \geq i)}. \tag{1}$$

Provided that our estimators Y are "good enough", we will have $\mathbb{E}[S] = 1/Z$. For example, Glynn et al. (2014) rely on the following lemma to show unbiasedness of their estimators:

Lemma 1. *$\mathbb{E}[S] = 1/Z$ if the following holds:*

$$\mathbb{E}\left[|Y^{(0)}| + \textstyle\sum_{i=1}^{\infty}|Y^{(i)} - Y^{(i-1)}|\right] < \infty. \tag{2}$$

The estimator in (1) is a Monte Carlo estimate of the infinite sum $Y^{(0)} + \sum_{i=1}^{\infty}(Y^{(i)} - Y^{(i-1)})$, which relies on the Y estimates becoming correct asymptotically. Condition (2) guarantees that the expectation of this Monte Carlo estimate is finite.

We can now define a baseline estimator inspired by Lyne et al. (2015). This estimator uses independent samples $X^{(0)}, \ldots, X^{(N)} \sim Q$ from which we set

$$Y^{(i)} = \frac{i+1}{\sum_{j=0}^{i} w(X^{(j)})}. \tag{3}$$

We will refer to this estimator as the Increasing Averages Estimator (IAE). Lyne et al. (2015) used a similar estimator, but with an exponentially increasing number of samples for each $Y^{(i)}$. So that we can make direct comparisons of individual design choices, all of the methods that we consider in this paper form estimates $Y^{(i)}$ based on a number of samples linear in i. Our proposed estimators work in this regime, and we could choose the distribution on N without worrying about running time growing out of control. However, our experiments are testing the individual theoretical proposals in this paper, not against the whole system that was originally proposed.

2.2 Pseudo-Marginal Markov Chain

We now review how to apply these unbiased estimates for $1/Z$ inside a pseudo-marginal outer MCMC loop. Recall that we have a class of densities $p(x\,|\,\theta) = p^*(x\,|\,\theta)/Z(\theta)$. Let $\pi(\theta)$ be a prior over the parameters, and y be a set of observations. Then the target posterior distribution is given by

$$\pi(\theta\,|\,y) \propto \frac{p^*(y\,|\,\theta)\,\pi(\theta)}{Z(\theta)}.$$

Standard Metropolis–Hastings sampling of this distribution, with proposal $t(\theta';\theta)$, computes the term

$$\min\left[1,\ \frac{p^*(y\,|\,\theta')\,\pi(\theta')\,t(\theta;\theta')\,Z(\theta)}{p^*(y\,|\,\theta)\,\pi(\theta)\,t(\theta';\theta)\,Z(\theta')}\right],$$

which requires the intractable ratio $Z(\theta)/Z(\theta')$.

A pseudo-marginal transition rule avoids needing to evaluate the normalizers exactly. Following the notation of Murray and Graham (2016), let $f(\theta) = p^*(y\,|\,\theta)\pi(\theta)/Z(\theta)$, with an unbiased estimate $\hat{f}$. If $\hat{f}$ is always positive, we can perform Metropolis–Hastings on the augmented state pair $(\theta, \hat{f})$. From the current state pair $(\theta, \hat{f})$, we propose a new state θ' with estimate $\hat{f}'$ and accept with probability

$$\min\left[1, \frac{\hat{f}'}{\hat{f}}\frac{t(\theta;\theta')}{t(\theta';\theta)}\right].$$

Unfortunately, the roulette estimator (1) can be negative if $Y^{(i)} - Y^{(i-1)} < 0$ for many values of i. Lyne et al. (2015) provide a clever way to avoid this "sign problem": replace the acceptance probability with

$$\min\left[1, \frac{|\hat{f}'|}{|\hat{f}|}\frac{t(\theta;\theta')}{t(\theta';\theta)}\right].$$

Then for each visited state $(\theta_i, \hat{f}_i)$, save σ_i, the sign of $\hat{f}_i$ such that $\hat{f}_i = \sigma_i|\hat{f}_i|$. Finally, when estimating the expectation of some function $h(\theta)$ over the posterior, the approximation $\sum_i h(\theta_i)\sigma_i / \sum_i \sigma_i$ is a consistent estimator for $\mathbb{E}_{\pi(\theta\,|\,y)}[h(\theta)]$.

A drawback to pseudo-marginal methods is that high variability in the estimator $\hat{f}(\theta)$ can encourage "sticking", as the same estimate $\hat{f}$ must be kept until a new state θ' is accepted. Furthermore, although the sign-normalized estimators are consistent, they will have high variance if a large fraction of the signs are negative. The construction of our Markov chain based estimators is motivated by the desire to address these issues.

3 Using a Markov Chain to Debias Importance Sampling Estimates

As motivation, we observe that the expectation of the inverse importance weights with respect to P is $1/Z$:

$$\mathbb{E}_{X\sim P}\left[\frac{1}{w(X)}\right] = \int \frac{P(X)Q(X)}{P^*(X)}\,\mathrm{d}X = \frac{1}{Z}.$$

Thus, samples drawn from P can provide unbiased estimates of $1/Z$. Although we can sample some target distributions P using coupling from the past (Propp and Wilson, 1998), for many choices of P no tractable exact sampling algorithm is known. However, using the tools from the previous section, we actually only need a sequence of samples whose distributions converge to P. We can obtain these with Markov chain Monte Carlo methods. We use the Metropolis–Hastings algorithm with proposals Q taken from an importance sampler.

We will use $(X = X^{(i)} : i \geq 0)$ to denote the states of our Markov chain. We can run a Markov chain whose stationary distribution converges to P as follows:

1. At time step i, draw a new state $X^{(i+1)}_{\text{prop}} \sim Q$.
2. Compute the acceptance ratio
$$a = \min\left[1, \frac{P(X^{(i+1)}_{\text{prop}})\,Q(X^{(i)})}{Q(X^{(i+1)}_{\text{prop}})\,P(X^{(i)})}\right] = \min\left[1, \frac{w(X^{(i+1)}_{\text{prop}})}{w(X^{(i)})}\right].$$
3. Draw a uniform random value $r^{(i)} \in [0,1]$ and set
$$X^{(i+1)} = \begin{cases} X^{(i)} & \text{if } r^{(i)} < a \\ X^{(i+1)}_{\text{prop}} & \text{if } r^{(i)} \geq a. \end{cases}$$

This chain and associated weights $w(X) = P^*(X)/Q(X)$ forms the backbone of our proposed debiasing schemes.

We need an asymptotically correct estimate Y. One obvious choice is $Y^{(i)} = 1/w(X^{(i)})$, where

$$\lim_{i\to\infty} \mathbb{E}[Y^{(i)}] = \lim_{i\to\infty} \mathbb{E}[1/w(X^{(i)})] = 1/Z,$$

since the distribution of $X^{(i)}$ approaches P. The main problem with this choice is that $\mathbb{E}[|Y^{(i)} - Y^{(i-1)}|]$ does not decay, as the chain might make large jumps from $X^{(i-1)}$ to $X^{(i)}$. As a result, the variance of the Russian roulette truncations will be high, and the final estimator might even have infinite expectation.

Glynn et al. (2014) suggests instead finding two sequences: $Y = (Y^{(i)} : i \geq 0)$ and $\tilde{Y} = (\tilde{Y}^{(i)} : i \geq 0)$ such that $\tilde{Y}^{(i)}$ follows the same distribution as $Y^{(i)}$, but $Y^{(i)}$ and $\tilde{Y}^{(i-1)}$ are likely to "couple" together. Then

$$S = Y^{(0)} + \textstyle\sum_{i=1}^{N} \frac{Y^{(i)} - \tilde{Y}^{(i-1)}}{\Pr(N \geq i)} \tag{4}$$

is an unbiased estimator of $1/Z$, since $Y^{(i)}$ and $\tilde{Y}^{(i)}$ follow the same distribution.

The pair of estimators $Y^{(i)}$ and $\tilde{Y}^{(i)}$ are constructed from Markov chains that share random numbers. Our Markov chain $X = (X^{(i)} : i \geq 0)$ uses a transition rule $\phi\colon \mathcal{X} \times \mathcal{X} \times [0,1] \to \mathcal{X}$, which uses a random number r to make each accept/reject decision:

$$X^{(i+1)} = \phi(X^{(i)}, X^{(i+1)}_{\text{prop}}, r^{(i)}),$$

where ϕ returns either the previous or proposed state according to the Metropolis–Hastings rule.

In what follows we write $\phi^{(i+1)}(\cdot) = \phi(\cdot, X^{(i+1)}_{\text{prop}}, r^{(i)})$ as the transition function determined by random choices of $X^{(i+1)}_{\text{prop}}$ and $r^{(i)}$. We also use $\tilde{X} = (\tilde{X}^{(i)} : i \geq 0)$ to denote a coupled copy of our chain. We would like to describe a coupling between X and $\tilde{X}$ so that $Y^{(i)} = 1/w(X^{(i)})$ and $\tilde{Y}^{(i)} = 1/w(\tilde{X}^{(i)})$ has the desired properties. We investigate alternative couplings in the following sections and defer formal guarantees of finite expectation to Section 4.

3.1 Forward Coupling

We use the following construction (Glynn et al., 2014):

$$\begin{aligned} X^{(i)} &= \phi^{(i)}(\phi^{(i-1)}(\ldots(\phi^{(1)}(X^{(0)})))) \\ \tilde{X}^{(i)} &= \phi^{(i+1)}(\phi^{(i)}(\ldots(\phi^{(2)}(X^{(0)})))) \end{aligned} \tag{5}$$

The chains $\tilde{X}^{(i)}$ and $X^{(i)}$ are dependent, and marginally come from the same distribution. Using $Y^{(i)} = 1/w(X^{(i)})$ and $\tilde{Y}^{(i)} = 1/w(\tilde{X}^{(i)})$ in (4) gives an unbiased estimate for $1/Z$. We can in fact compute $Y^{(i)}$ and $\tilde{Y}^{(i)}$ only knowing the sequence of proposed weights $(w(X^{(i)}_{\text{prop}}) : 0 \leq i \leq N)$ without requiring exact knowledge of the states $\tilde{X}^{(i)}$. This makes it simple to implement our debiasing scheme given access to a black-box importance sampler. We refer to this estimator as the forward coupled estimator (FCE) and illustrate it concretely in Algorithm 1.

The key feature of our estimator is that it attempts to couple together $X^{(i)}$ and $\tilde{X}^{(i-1)}$ by subjecting both chains to the same sequence of random transitions following $\phi^{(1)}$, which is applied to X but not $\tilde{X}$. If $X^{(i-1)}$ and $\tilde{X}^{(i-2)}$ both accept $X^{(i)}_{\text{prop}}$ when subjected to $\phi^{(i)}$, then $X^{(i)} = \tilde{X}^{(i-1)}$, and X and $\tilde{X}$ couple together. All the subsequent correction terms cancel out if $X^{(i)}$ and $\tilde{X}^{(i-1)}$ have coupled: in (4), $Y^{(j)} - \tilde{Y}^{(j-1)} = 0$ for all $i \leq j \leq N$. This cancellation serves as a form of variance reduction for our estimator.

Algorithm 1 Forward Coupled Estimator

Input: Target distribution P and proposal distribution Q.
Output: S, an unbiased estimate for $1/Z$

Draw random stopping time N.
Draw $X^{(0)}_{\text{prop}}, \dots, X^{(N)}_{\text{prop}} \sim Q$ and initialize $w^{(0)}, \dots, w^{(N)}$ with $w^{(i)} = w(X^{(i)}_{\text{prop}})$.
Initialize $S = 1/w^{(0)}$, $w = w^{(0)}$, $\tilde{w} = w^{(0)}$.
for $i = 1$ to N **do**
  Draw $r^{(i-1)} \sim \text{Uniform}[0, 1]$.
  Compute $a = \min\{1, w^{(i)}/w\}$.
  Compute $\tilde{a} = \min\{1, w^{(i)}/\tilde{w}\}$.
  if $r^{(i-1)} < a$ **then**
    Update $w = w^{(i)}$.
  end if
  if $r^{(i-1)} < \tilde{a}$ and $i > 1$ **then**
    Update $\tilde{w} = w^{(i)}$.
  end if
  Update $S = S + \frac{w^{-1} - \tilde{w}^{-1}}{\Pr(N \geq i)}$.
end for

We can provide a simple lower bound on the probability of coupling by time step i, which also translates into a method for guaranteeing positive estimates.

Lemma 2. *For the FCE, if $i \geq 2$,*

$$\Pr[X^{(i)} \text{ and } \tilde{X}^{(i-1)} \text{ have coupled }] \geq 1 - \tfrac{2}{i+1}.$$

Proof. Let j^* be the smallest $2 \leq j \leq i$ such that $w(X^{(j)}_{\text{prop}}) \geq \max\{w(X^{(0)}), w(X^{(1)}_{\text{prop}})\}$, if such a j exists. Then both chains X and $\tilde{X}$ must accept the proposal at $\phi^{(j^*)}$ since $w(X^{(j^*)}_{\text{prop}}) > \max\{w(X^{(j^*-1)}), w(\tilde{X}^{(j^*-2)})\}$ so the acceptance ratios evaluate to 1. The probability of j^* existing is at least $\frac{i-1}{i+1}$, the probability that the largest importance weight is proposed between the second and i-th proposal since our importance weights are drawn i.i.d. Thus, the two chains would have coupled with probability at least $1 - 2/(i+1)$. □

If the Markov chain estimator discarded an initial "burn-in" period of T time steps, we can guarantee that our estimates will have a probability of at least $1 - 1/T$ of being positive after debiasing. Concretely, define Y and $\tilde{Y}$ alternatively so that $Y^{(i)} = 1/w(X^{(i+T)})$ and $\tilde{Y}^{(i)} = 1/w(\tilde{X}^{(i+T)})$. Then Lemma 2 implies the following result:

Proposition 1. *Compute S as in Algorithm 1, except allowing for the burn-in of T steps. Then*

$$\Pr[S \geq 0] \geq 1 - \tfrac{2}{T+1}$$

This result follows simply from noting that if $X^{(T)}$ and $\tilde{X}^{(T-1)}$ are coupled, then $Y^{(i)} - \tilde{Y}^{(i-1)} = 0$ for $i \geq 1$. In fact, a simple argument can improve the probability to $1 - 1/(T+1)$, which we omit for space reasons. Our experiments did not use a burn-in period for ease of comparison. Even without burn-in, FCE gives a higher percentage of positive estimates than other formulations.

FCE can have high variance when the underlying importance sampler is variable. If $X^{(1)}_{\text{prop}}$ is very large, coupling may be impeded because $\tilde{X}$ does not encounter this proposal, and X will have difficulty moving away from $X^{(1)}_{\text{prop}}$ due to low acceptance probabilities. Our next estimator improves this situation, although provides fewer guarantees on positive estimates.

3.2 Backward Coupling

We use an alternative construction for X and $\tilde{X}$, also from Glynn et al. (2014):

$$X^{(i)} = \phi^{(N)}(\phi^{(N-1)}(\dots(\phi^{(N-i+1)}(X^{(N-i)}_{\text{prop}})))), \quad (6)$$

where N is the random stopping time of the Russian roulette truncation. We refer to this coupling as "backwards" because we process the proposals in reverse. For this estimator, we will simply let $\tilde{X}^{(i)} = X^{(i)}$.

We will also reduce the variance of our estimates by computing the expectations of $1/w(X^{(i)})$ over the random draws r used to determine acceptance. The process of averaging out r is a case of a general technique called Rao–Blackwellization, which has been shown to reduce variance when applied to Metropolis–Hastings sampling updates (Casella and Robert, 1996). We can formally express $Y^{(i)}$ as follows: first independently sample proposals $X^{(0)}_{\text{prop}}, \dots, X^{(N)}_{\text{prop}} \sim Q$. Then

$$Y^{(i)} = \mathbb{E}\left[\frac{1}{w(\phi^{(N)}(\dots(X^{(N-i)}_{\text{prop}})))} | X^{(N)}_{\text{prop}}, \dots, X^{(N-i)}_{\text{prop}}\right]. \quad (7)$$

Since $X^{(N)}_{\text{prop}}, \dots, X^{(N-i)}_{\text{prop}}$ are given, this equation denotes the expectation of $1/w(X^{(i)})$ with $r^{(N-1)}, \dots, r^{(N-i)}$ averaged out. By the law of iterated expectations, we still have $\lim_{i\to\infty} \mathbb{E}[Y^{(i)}] = 1/Z$, so our Rao–Blackwellized estimator is unbiased in $1/Z$.

Example 1. *In the case where $i = 1$,*

$$Y^{(1)} = \frac{1}{w(X^{(N-1)}_{prop})}\left(1 - \min\left[1, \frac{w(X^{(N)}_{prop})}{w(X^{(N-1)}_{prop})}\right]\right) + \frac{1}{w(X^{(N)}_{prop})} \min\left[1, \frac{w(X^{(N)}_{prop})}{w(X^{(N-1)}_{prop})}\right]$$

We outline the Rao–Blackwellization process in Algorithm 2. We refer to our estimator as the Rao–Blackwellized backward coupled estimator (RBBCE). Like FCE, RBBCE only requires knowledge of the importance weights, not the states, to run. The algorithm is simple to implement and provably fast in expectation. The following proposition shows that we can perform Rao–Blackwellization essentially "for free" on the backward coupled estimator.

Proposition 2. *Algorithm 2 takes expected $O(N)$ running time.*

Proof. Following the notation in Algorithm 2, we will let $w^{(i)} = w(X^{(i)}_{\text{prop}})$. We can compute the expected

Algorithm 2 Rao–Blackwellized Backward Coupled Estimator

Input: Target distribution P and proposal distribution Q.
Output: S, an unbiased estimate for $1/Z$

1: Draw random stopping time N.
2: Draw $X^{(0)}_{\text{prop}}, \dots, X^{(N)}_{\text{prop}} \sim Q$ Initialize $w^{(0)}, \dots, w^{(N)}$ with $w^{(i)} = w(X^{(i)}_{\text{prop}})$.
3: Initialize $S = 1/w^{(N)}$, $Y^{(0)}_{\text{rb}} = 1/w^{(N)}$.
4: **for** $i = 1$ to N **do**
5: Find k, $0 \le k < i$ such that $w^{(N-k)} = \max_{0 \le j < i} w^{(N-j)}$.
6: **if** $w^{(N-i)} < w^{(N-k)}$ **then**
7: Set $Y^{(i)}_{\text{rb}} = Y^{(k)}_{\text{rb}}$.
8: **else**
9: Initialize $Y^{(i)}_{\text{rb}} = 0$, $\gamma = 1$.
10: **for** $j = 0$ to $i - 1$ **do**
11: Update

$$Y^{(i)}_{\text{rb}} = Y^{(i)}_{\text{rb}} + \frac{w^{(N-i+j+1)}}{w^{(N-i)}} \cdot \gamma \cdot Y^{(i-j-1)}_{\text{rb}}$$

12: Update

$$\gamma = \gamma \cdot \left(1 - \frac{w^{(N-i+j+1)}}{w^{(N-i)}}\right)$$

13: **end for**
14: Update $Y^{(i)}_{\text{rb}} = Y^{(i)}_{\text{rb}} + \gamma \cdot \frac{1}{w^{(N-i)}}$.
15: **end if**
16: Update $S = S + \frac{Y^{(i)}_{\text{rb}} - Y^{(i-1)}_{\text{rb}}}{\Pr(N \ge i)}$.
17: **end for**

runtime of each iteration of the loop at line 4. If the current proposed weight at iteration i, $w^{(N-i)}$, is less than $\max_{0 \le j < i} w^{(N-j)}$, then the chain will always accept at this maximum because the acceptance ratio will be 1. In this case, we take $O(1)$ time to update $Y^{(i)}_{\text{rb}}$. If the current proposed weight is greater than $\max_{0 \le j < i} w^{(N-j)}$, then we take $O(i)$ time to compute $Y^{(i)}_{\text{rb}}$. The probability of this happening is $\frac{1}{i+1}$ because $w^{(N-i)}, \dots, w^{(N)}$ are i.i.d. draws, so the total expected runtime of each iteration is $O(1) + O(i/(i+1)) = O(1)$. The loop runs N times giving expected $O(N)$ runtime. □

The $O(N)$ expected time means Rao–Blackwellization only adds a constant cost to the computation of each importance weight, which will be negligible for expensive, low-variance weights.

We can explain "coupling" in RBBCE as follows: $Y^{(i)} = Y^{(i-1)}$ unless $w(X^{(N-i)}_{\text{prop}}) > \max_{N-i+1 \le j \le N} w(X^{(j)}_{\text{prop}})$, because otherwise the acceptance probability will be 1. Thus, $Y^{(i)} - Y^{(i-1)}$ only has probability $1/(i+1)$ of being nonzero. In comparison, the difference terms in IAE contribute to higher variance because they are only nonzero if

$$w(X^{(i)}) = \left[\sum_{j=0}^{i-1} w(X^{(j)})\right] \Big/ i,$$

which occurs with extremely low probability. We find empirically that RBBCE obtains lower variance.

3.3 Averaging batches of importance weights

Taking the reciprocal of importance weights in Algorithm 1 or 2 will give high variance estimates if the weights are occasionally small. We reduce the variance of the importance weights by averaging over a batch:

$$w = \frac{1}{m} \sum_{i=1}^{m} \frac{P^*(X_i)}{Q(X_i)}. \tag{8}$$

One way to justify using these average weights in the Markov chains is to define new targets and proposals P_m and Q_m on the augmented state space $\mathcal{X}^m$:

$$Q_m(X_1, \dots, X_m) = \textstyle\prod_{i=1}^m Q(X_i),$$
$$P^*_m(X_1, \dots, X_m) = \tfrac{1}{m} \textstyle\sum_{i=1}^m P^*(X_i) \prod_{j \ne i} Q(X_j).$$

Because Q is normalized, it follows that the normalizer for P^*_m is Z. Using P_m and Q_m as target and proposal distributions means that the weights in Algorithms 1 and 2 become the average of a batch of weights (8).

4 Unbiasedness in $1/Z$

Now we will formally establish the unbiasedness properties of our proposed estimators. First, we will formally define when the expectation of a random variable is finite. Our motivation is to characterize when the Law of Large Numbers (LLN) holds for FCE and RBBCE.

Definition 1. *Let A be a random variable with state space $\mathcal{A}$. Let $f : \mathcal{A} \to \mathbb{R}$ be a real-valued function, and let λ be the distribution of A on $\mathcal{A}$. Then we say that A has finite expectation if $\int_A |f(A)| d\lambda < \infty$ and A has infinite expectation otherwise.*

The distinction between finite and infinite expectation is important because the LLN only applies to random variables with finite expectation. We rely on Lemma 1 to show that FCE has finite expectation whenever $\mathcal{X}$ is a finite state space. We also show that RBBCE always has finite expectation for any choice of state space $\mathcal{X}$.

Proposition 3. *Let Q and P have full support over $\mathcal{X}$. So long as $\mathcal{X}$ is finite, the output of Algorithm 1 will have finite expectation and so is unbiased in $1/Z$.*

Proof. Since $\mathcal{X}$ is finite and Q and P have full support, we can define the maximum and minimum possible importance weights by

$$w_{\min} = \min_{X \in \mathcal{X}} w(X), w_{\max} = \max_{X \in \mathcal{X}} w(X)$$

Define $X_{\min}$ and $X_{\max}$ as states corresponding to $w_{\min}$ and $w_{\max}$. Now recall that $Y^{(i)} = 1/w(X^{(i)})$ and $\tilde{Y}^{(i-1)} = 1/w(\tilde{X}^{(i-1)})$. If $X_{\max}$ is proposed by $\phi^{(j)}$ for $2 \le j \le i$, then both X and $\tilde{X}$ must accept at $\phi^{(j)}$ with probability 1. In this case, $Y^{(i)} - \tilde{Y}^{(i-1)} = 0$. Now if this does not happen, then the trivial upper bound

$$|Y^{(i)} - \tilde{Y}^{(i-1)}| \le \left|\tfrac{1}{w_{\min}} - \tfrac{1}{w_{\max}}\right|$$

must apply. We can upper bound the probability that $X_{\max}$ is not proposed by $(1 - Q(X_{\max}))^{i-1}$ since proposals are drawn independently. For $i \geq 2$, this gives us an expected value bound

$$\mathbb{E}[|Y^{(i)} - \tilde{Y}^{(i-1)}|] \leq (1 - Q(X_{\max}))^{i-1} \left| \tfrac{1}{w_{\min}} - \tfrac{1}{w_{\max}} \right|$$

and therefore

$$\mathbb{E}\left[|Y^{(0)}| + \textstyle\sum_{i=1}^{\infty} |Y^{(i)} - \tilde{Y}^{(i-1)}|\right] \leq \mathbb{E}[|Y^{(0)}|] + \mathbb{E}[|Y^{(1)} - \tilde{Y}^{(0)}|] + \textstyle\sum_{i=2}^{\infty}(1 - Q(X_{\max}))^{i-1} \left| \tfrac{1}{w_{\min}} - \tfrac{1}{w_{\max}} \right| < \infty,$$

because $(1 - Q(X_{\max})) < 1$ since Q has full support on $\mathcal{X}$, and therefore the equation is a geometric series. Thus, (2) is satisfied (we note that this is for $Y^{(i)} - \tilde{Y}^{(i-1)}$ instead of $Y^{(i)} - Y^{(i-1)}$, but Lemma 1 still applies) so Lemma 1 completes the proof. □

For RBBCE, we can provide even stronger guarantees for unbiasedness. In particular, even if $\mathcal{X}$ is infinite, RBBCE will always have finite expectation so long as Q and P have full support over $\mathcal{X}$ and

$$\mathbb{E}_{X \sim Q}[Q(X)/P^*(X)] < \infty. \tag{9}$$

This assumption ensures that $\mathbb{E}[Y^{(0)}] < \infty$ and is a natural assumption to make for reasonable choices of P. To prove our result, we require the following observation about $Y^{(i)}$:

Lemma 3. *Recall that for RBBCE, $Y^{(i)}$ is defined in (7). For any $i \geq 1$, $Y^{(i)} \leq Y^{(i-1)}$.*

Proof. For $Y^{(i)}$, recall that the Markov chain first starts at state $X_{\text{prop}}^{(N-i)}$. We will first analyze what happens for each fixed choice of $\phi^{(N)}, \ldots, \phi^{(N-i)}$ and then average out the random draws $r^{(N-1)}, \ldots, r^{(N-i)}$. First, let

$$\hat{Y^{(i)}} = 1/w(\phi^{(N)}(\ldots(X_{\text{prop}}^{(N-i)}))).$$

$\hat{Y^{(i)}}$ denotes an instantiation of $Y^{(i)}$ without Rao–Blackwellization over random acceptances. Let $k = \max_{0 \leq j < i} w(X_{\text{prop}}^{(N-j)})$. There are two cases for $X_{\text{prop}}^{(N-i)}$: if $w(X_{\text{prop}}^{(N-i)}) \leq w(X_{\text{prop}}^{(N-k)})$, then both the chains for $\hat{Y^{(i)}}$ and $\hat{Y^{(i-1)}}$ must accept $X_{\text{prop}}^{(N-k)}$, in which case $\hat{Y^{(i)}} = \hat{Y^{(i-1)}}$. When $w(X_{\text{prop}}^{(N-i)}) > w(X_{\text{prop}}^{(N-k)})$, consider the first acceptance by the chain for $\hat{Y^{(i)}}$. Since $w(X_{\text{prop}}^{(N-i)}) > w(X_{\text{prop}}^{(N-k)})$, the corresponding acceptance ratio for $\hat{Y^{(i-1)}}$ is greater than the acceptance ratio for $\hat{Y^{(i)}}$ at this point. Thus, the chain for $\hat{Y^{(i-1)}}$ must also accept at this point, resulting in coupling, so $\hat{Y^{(i)}} = \hat{Y^{(i-1)}}$ again. Finally, if the chain for $\hat{Y^{(i)}}$ never accepts, then

$$\hat{Y^{(i)}} = \frac{1}{w(X_{\text{prop}}^{(N-i)})} < \frac{1}{w(X_{\text{prop}}^{(N-k)})} \leq \hat{Y^{(i-1)}}.$$

In all cases, $\hat{Y^{(i)}} \leq \hat{Y^{(i-1)}}$. Thus,

$$Y^{(i)} - Y^{(i-1)} = \mathbb{E}[\hat{Y^{(i)}} - \hat{Y^{(i-1)}} | X_{\text{prop}}^{(N)}, \ldots, X_{\text{prop}}^{(N-i)}] \leq 0$$

□

Proposition 4. *As long as Q and P have full support over $\mathcal{X}$ and (9) holds, the output of Algorithm 2 will be unbiased in $1/Z$ and have finite expectation.*

Proof. From Lemma 3, it follows that $Y^{(i)} - Y^{(i-1)} \leq 0$ so $\mathbb{E}[|Y^{(i)} - Y^{(i-1)}|] = -\mathbb{E}[Y^{(i)} - Y^{(i-1)}]$ for $i \geq 1$. Therefore,

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}[|Y^{(i)} - Y^{(i-1)}|] &= -\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)} - Y^{(i-1)}] \\ &= -\lim_{n\to\infty} (\mathbb{E}[Y^{(n)}] - \mathbb{E}[Y^{(0)}]) \\ &= -1/Z + \mathbb{E}[Y^{(0)}] < \infty.\end{aligned}$$

It follows that (2) holds, so Lemma 1 implies that Algorithm 2 provides an output unbiased in $1/Z$. □

4.1 Comparison To Existing Russian Roulette Estimator

We do not know of any proofs of finite expectation for the IAE estimator described in (3). A simple example shows that (2) can be violated.

Example 2. *Consider the case where $\mathcal{X} = \{0, 1\}$, $Q(0) = Q(1) = 1/2$, and $P^*(0) = 1$, $P^*(1) = 2$. Then if we define $Y^{(i)}$ as in (3),*

$$\mathbb{E}\left[|Y^{(0)}| + \sum_{i=1}^{\infty} |Y^{(i)} - Y^{(i-1)}|\right]$$

is infinite. In particular, (2) is not satisfied.

Explanation for claim. We show that $\mathbb{E}[|Y^{(i)} - Y^{(i-1)}|] = \Omega(1/i)$. Consider the event E_i where at least half of the proposed states X_j for $j < i$ are 0, and $X_i = 1$. Let $S_i = \sum_{j=0}^{i} w(X_i)$.

$$Y^{(i)} - Y^{(i-1)} = \frac{i+1}{S_i} - \frac{i}{S_i - 4} = \frac{S_i - 4i - 4}{S_i(S_i - 4)}$$

Now since over half the states X_j with $j < i$ are 0, it follows that $S_i \leq 3i + 4$. Thus,

$$Y^{(i)} - Y^{(i-1)} \leq -\frac{1}{9i + 12}.$$

So with probability at least $\Pr[E_i]$, $|Y^{(i)} - Y^{(i-1)}| \geq \frac{1}{9i+12}$. From inspecting Q, it is evident that $\Pr[E_i] \geq 1/4$, so $\mathbb{E}[|Y^{(i)} - Y^{(i-1)}|] \geq \frac{1}{36i+48}$. Summing over all i gives a divergent infinite sum. □

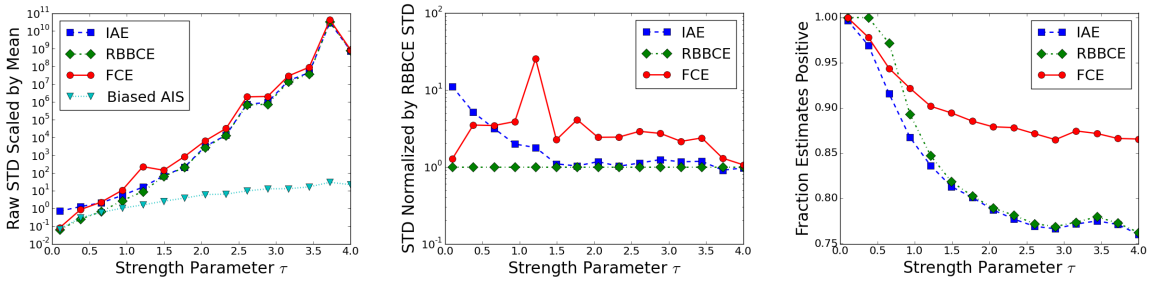


Figure 1: $1/Z$ estimator performance for Ising models with different values of τ. Each estimator is run for 10,000 trials. **Left:** Standard deviation divided by the mean of the estimator. For IAE, RBBCE, and FCE, this is $1/Z$, which we know exactly. The biased AIS estimator is the inverse importance weights, and we plot empirical standard deviation over empirical mean. **Center:** Each standard deviation is divided by the RBBCE standard deviation, for clearer comparison. **Right:** The fraction of positive estimates returned by each estimator.

Our analysis highlights an advantage of Markov chain based estimators: without the need for a case-by-case analysis or for tuning $Y^{(i)}$ to require a superlinear number of samples in i, our estimators are guaranteed to have the correct expectation for many choices of $\mathcal{X}$ (all choices in the case of RBBCE).

5 Demonstrations

We test empirically how the estimators work in practice. Following Møller et al. (2006), we test our algorithms on a grid Ising model, a graphical model with nodes I and edges E parametrized by

$$p(x \mid \alpha, \beta) = \frac{1}{Z(\alpha,\beta)} \left(\sum_i \alpha_i + \sum_{i \neq j \in E} \beta_{ij} x_i x_j \right).$$

For our Ising model, we use a 10×30 lattice graph. In each experiment we set a strength parameter τ, and randomly sampled each α_i and β_{ij} from Uniform$[-\tau, \tau]$.

We estimated the standard deviations of the $1/Z$ estimators by computing the empirical root mean square error from the true value. It is possible to compute $Z(\alpha, \beta)$ exactly for the narrow strip we used. We also evaluated the empirical fraction of positive estimates for each algorithm.

Our importance sampling estimates were based on AIS (Neal, 2001) using 10 intermediate distributions. We used the averaging scheme described in Section 3.3 and average over 10 AIS weights before taking reciprocals, which significantly improved variance. For our distribution on N, we choose the distribution satisfying $\Pr(N \geq k) \propto k^{-1.1}$.

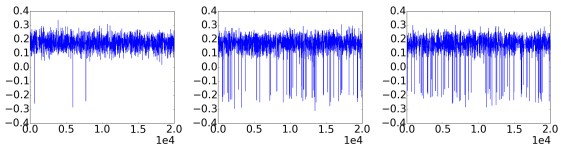
Figure 1 shows the Markov-chain based estimators have lower variance and more positive estimates than IAE for lower values of τ, where the importance samplers work well. We show the variance of the inverse importance weights as a reference to show how much debiasing increases variance. The variance of all three estimators increases as the importance sampling estimates become less reliable, but FCE degrades fastest because the Markov chain within FCE is more likely to "stick". However, FCE retains a significantly higher percentage of positive estimates, as expected theoretically.

At higher values of τ, where the importance sampling estimates are less reliable, the IAE and RBBCE curves begin to look more similar. RBBCE still outperforms IAE in both variance and percent positive estimates for almost all values of τ. In practice, however, it would make sense to improve the importance sampling estimates by increasing the number of intermediate annealing distributions and averaging over more estimates before applying debiasing schemes. In the setting where importance sampling estimates are already reliable, our Markov chain based estimators perform much better.

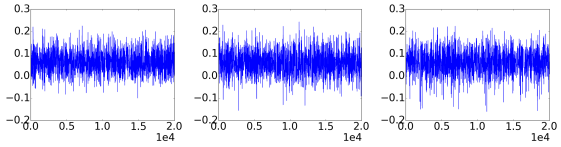
5.1 Pseudo-marginal Ising Grid

We next tested the Russian Roulette algorithms in a pseudo-marginal estimation setting. We again run our experiments on a 10×30 Ising lattice. We use a single bias and coupling parameter: $\alpha_i = \alpha$ and $\beta_{ij} = \beta$. Following Murray and Graham (2016), we use uniform priors over $\alpha \in [-1, 1]$ and $\beta \in [0, 0.4]$. We used data generated with $\alpha = 0.1$ and $\beta = 0.1$. The pseudo-marginal Metropolis–Hastings outer loop used Gaussian proposals: $\alpha' \sim \mathcal{N}(\alpha, 0.025^2)$ and $\beta' \sim \mathcal{N}(\beta, 0.01^2)$, was run for 100,000 iterations, and used the method of Lyne et al. (2015) for dealing with negative estimates. Our unbiased $1/Z$ estimator was averaged over 2 trials, each trial used weights formed by averaging 10 AIS weights with 30 intermediate distributions.

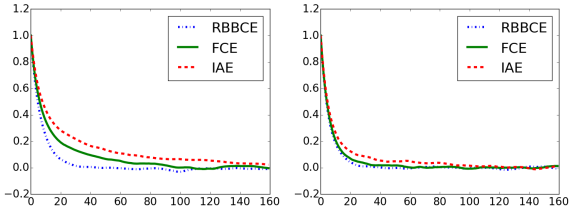
Figure 2 shows the empirical autocorrelations and trace plots for our experiments. As $\beta > 0$ negative values in the trace plot indicates a negative estimate for $1/Z$. Overall, out of 100,000 iterations, RBBCE had 99,924 positive samples while FCE and IAE had 97,597 and 96,538, respectively. As discussed by Lyne et al. (2015), a large fraction of positive estimates gives lower variance estimates of the posterior, which favours RBBCE.



(a) Trace plots for $\sigma\beta$. From left to right: RBBCE, FCE, IAE



(b) Trace plots for $\sigma\alpha$. From left to right: RBBCE, FCE, IAE



(c) Autocorrelation vs. lag for $\sigma\beta$ (left) and $\sigma\alpha$ (right).

Figure 2: Trace and autocorrelation plots for doubly-intractable Ising runs. All plots tracked parameters multiplied by σ, the sign of the estimator for $1/Z$. The autocorrelations without the sign term are roughly the same for all methods. Negative values in (a) result from negative σ, which gives high variance estimates.

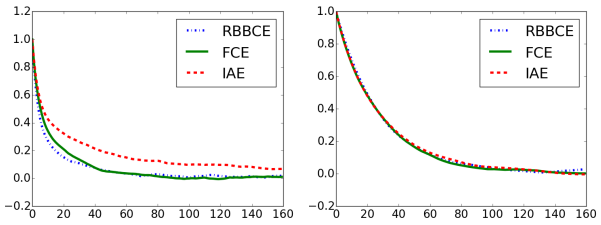


Figure 3: Autocorrelation vs. lag for θ_e (left) and θ_s (right). As in Fig. 2(c), we plot signed autocorrelations. Markov chain based estimators exhibit less sticking.

We have not compared to the exchange algorithm (Murray et al., 2006), which applies to this specific Ising model example. A direct comparison would be difficult: unlike our methods, the exchange algorithm depends on exact sampling, which has highly variable cost and depends on several additional details.

5.2 Exponential Random Graph Model

In our final demonstration, we apply the pseudo-marginal chains to Bayesian inference on exponential random graph models (Caimo and Friel, 2011). These models capture relationships between sets of nodes, such as social interactions between individuals or formation of chemical structures between atoms. The distribution over graphs is

$$p(x \mid \theta) = \exp(\theta^T s(x))/Z(\theta),$$

where θ are parameters and s is a vector of sufficient statistics of the graph x. Caimo and Friel (2011) used the exchange algorithm (Murray et al., 2006) with approximate rather than exact samples. We believe our experiments are the first application of an asymptotically correct MCMC method to these models.

Our experiments use the Florentine graph, a social network graph modeling business relations between families in Florence in 1430. We let $\theta = (\theta_e, \theta_s)$, and $s = (\text{number of edges}, \text{average number of 2-stars per node})$, where a node with degree d is involved in $\binom{d}{2}$ 2-stars. We use a uniform prior for θ_e on $[-2.5, 2.5]$ and a uniform prior for θ_s on $[-1, 1]$. We run our pseudo-marginal chains for 100,000 iterations, averaging over 10 trials for each unbiased $1/Z$ estimator and using averages of 10 AIS weights with 10 intermediate distributions for our importance sampler. We tune Gaussian steps to 1 for θ_e and 0.1 for θ_s.

Figure 3 shows the empirical autocorrelations of our chains. We report 99,890 positive estimates for RBBCE, 98,680 for FCE, and 98,442 for IAE. Although the improvements in positive estimates are more modest this time, our Markov chain based estimators still demonstrate lower autocorrelations than IAE.

6 Discussion

We introduced two novel algorithms, FCE and RBBCE, for producing unbiased estimates of $1/Z$ given access to black-box estimates unbiased in Z. Our algorithms are generic, simple to implement, and perform debiasing at virtually no added cost. We are able to provide theoretical guarantees of finite expectation for many choices of state space (all choices for RBBCE) that hold regardless of the underlying distribution on truncation time. Unlike existing methods, these results allow valid use of the algorithms without needing to tune free parameters such as the growth rate of number of importance samples with truncation time.

FCE and RBBCE rely on Markov chain "coupling" with the motivation of improving variance and percentage of positive estimates, two heuristic indicators for how well our estimators would perform in a pseudo-marginal outer loop. Our experiments demonstrate that our algorithms can provide promising improvements over a non-coupling based debiasing scheme.

Our debiasing framework could be freely combined with recent developments in pseudo-marginal MCMC. For example Doucet et al. (2015)'s analysis could be used to tune the number of samples used for the $1/Z$ estimate. We could also apply pseudo-marginal slice sampling (Murray and Graham, 2016) with our algorithms.

References

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697–725, 2009.

A. Caimo and N. Friel. Bayesian inference for exponential random graph models. *Social Networks*, 33 (1):41–55, 2011.

L. L. Carter and E. Cashwell. Particle-transport simulation with the Monte Carlo method. Technical report, Los Alamos Scientific Lab., 1975.

G. Casella and C. P. Robert. Rao–Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, page asu075, 2015.

P. W. Glynn, C.-h. Rhee, et al. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51:377–389, 2014.

P. E. Jacob, A. H. Thiery, et al. On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2): 769–784, 2015.

I. Lux and L. Koblinger. *Monte Carlo particle transport methods: neutron and photon calculations*, volume 102. CRC press, 1991.

A.-M. Lyne, M. Girolami, Y. Atchade, H. Strathmann, D. Simpson, et al. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467, 2015.

D. McLeish et al. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications*, 17(4):301–315, 2011.

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

I. Murray and M. M. Graham. Pseudo-marginal slice sampling. *JMLR: W&CP*, 51:911–919, 2016.

I. Murray, Z. Ghahramani, and D. J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2006.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

J. Propp and D. Wilson. Coupling from the past: a user's guide. *Microsurveys in Discrete Probability*, 41:181–192, 1998.